Gudrun Rawoens

Korpuslingvistik och kontrastiv språkbeskrivning. Ett svensk-nederländskt projekt.

(publiceras 2003) Korpuslingvistik och kontrastiv språkbeskrivning. Ett svensk-nederländskt projekt. I: *Tijdschrift voor Scandinavistiek*.

Inledning

I artikelns första del ämnar jag belysa begreppen korpuslingvistik och kontrastiv språkforskning och med hjälp av några exempel påvisa de olika användningsområdena i dagens språkstudier.

I andra delen kommer jag att beskriva det konkreta arbetet med uppbyggnaden av den svensk-nederländska parallellkorpusen som jag använder mig av inom ramen för mitt avhandlingsarbete om kausativitet i svensk-nederländskt kontrastivt perspektiv.

Korpuslingvistik

Definition

Aijmer och Altenberg (1996:1) definierar korpuslingvistik som ”the study of language on the basis of text corpora”. Enligt Leech (1992:105) visar begreppet inte till ”a domain of study, but rather to a methodological basis for pursuing linguistic research”. Termen korpuslingvistik åsyftar en heuristisk metod som används inom språkforskning. Leech (ibidem:107) nämner fyra områden som korpuslingvistiken fokuserar på: språklig performans i stället för kompetens, språkbeskrivning i stället för språkliga universalia, både kvantitativa och kvalitativa språkmodeller, och infallsvinkeln är empirisk snarare än rationell (”*rationalist*”).

McEnery & Wilson (2001) ägnar ett helt verk åt att förklara begreppet korpuslingvistik, men sammanfattar det i början av boken på ett enkelt sätt:

> Corpus linguistics is perhaps best described for the moment in simple terms as the study of language based on examples of ‘real life’ language use. (McEnery & Wilson 2001:1)

Ooi (1998:34) hänvisar till Ooi (1994:2) och säger att en korpuslingvistisk studie bedrivs ”on the basis of textual or acoustic corpora, almost always involving the computer in some phase of storage, processing, and analysis of this data.”

Vad är en korpus?

I språkforskning som bedrivs inom korpuslingvistiken utgör en korpus alltså den empiriska basen. En korpus kan definieras som en samling språkliga yttranden som någon gång har blivit nedskrivna (skriftkorpus) eller sagda (talkorpus). Den enkla definitionen som McEnery & Wilson (2001:29) ger är ”any collection of more than one text”. De fortsätter med en utförlig förklaring och uppger fyra krav som en korpus ska uppfylla: en korpus ska ha

en bestämd representativitet, den ska vara begränsad, datalagrad och fungera som standardreferens till det språk den representerar. Uppfylls inte de här kriterierna bör man snarare tala om ”a collection of texts” i stället (McEnery & Wilson 2001:103). Vad dessa kriterier innebär kommer jag tillbaka till.

Historik

Korpusar har i själva verket funnits lika länge som boktryckarkonsten. Tidiga språkforskare[1] använde sig av korpusar för att studera en översättning eller ett bestämt språkligt fenomen.

Användningen av korpuslingvistiken som heuristisk metod såsom den tillämpas i dagens språkstudier, började få sin nuvarande form först från och med 1960-talet efter ett uppehåll under 50-talet, då många lingvister till följd av Chomskys stora inflytande och hans starka kritik mot empiriska metoder avstod från att använda korpusar[2] (se bl.a. McEnery & Wilson 2001:5-10). Under en period där den systemorienterade lingvistiken var så pass dominerande var det helt enkelt omöjligt för den språkbruksorienterade metoden att slå igenom.

På 60-talet, men framförallt från och med 70-talet fick metoden ny uppmärksamhet och snabbt många anhängare. Detta hängde också ihop med datorns tillkomst som möjliggjorde att man kunde samla in stora mängder textmaterial i elektronisk form (se bl.a. McEnery & Wilson 2001:17 ; Leech 1992:105). Under det senare 60-talet påbörjades insamlingen av textkorpusar över engelska språket. The Brown University Corpus of American English[3] (Francis & Kucera 1967), och Lancaster-Oslo/Bergen Corpus (LOB)[4] (Hofland & Johansson 1982) tillhör de första datalagrade korpusarna.

I och med att man i större skala började använda sig av datakorpusar etablerades termen *computer corpus linguistics* eller *datakorpuslingvistik*, vanligtvis kallad *korpuslingvistik*, som beteckning för metoden. Talar man om dagens korpusar, tänker man endast på datorlagrade korpusar, vilket har medfört att termen *korpuslingvistik* har blivit den gängse benämningen[5]. I takt med datautvecklingen har det också blivit vanligare med allt större korpusar som kan innehålla flera miljoner ord[6]. De första korpusarna var enspråkiga korpusar. Det senaste decenniet har det också blivit vanligt med två- eller flerspråkiga korpusar. Intresset för flerspråkskorpusar växte från och med 90-talet i takt med det växande intresset för kontrastiv språkforskning och språkteknologi.

[1] Svartvik (1992:7) citerar Otto Jespersen (1938) som beskriver hur han under många år samlade in autentiska språkexempel som han senare återanvände i sin forskning.
[2] Chomskys kritik gick ut på att han ansåg att språkforskaren måste studera den språkliga *kompetensen* (den abstrakta språkförmågan som varje modersmålstalare har) snarare än *performansen* (språkbeteendet). För lingvistiskt studium var kompetensen enligt Chomsky bara åtkomlig genom modersmålstalarens/lingvistens egen intuition. Eftersom performansen som uttryck för verkligt språkbeteende påverkas av många andra faktorer (t.ex. situationsbundna) kan den därför inte återspegla kompetensen, menade Chomsky. Även på senare år har Chomsky betonat klyftan mellan ”the E-language (”externalised language”)” eller den språkliga performansen å ena sidan, och ”the I-language (”internalised language”) eller den språkliga kompetensen å andra sidan. (Cook 1988:12) Studieobjektet i korpuslingvistiken är alltså E-språket.
[3] The Brown Corpus sammanställdes av W. Nelson Francis och Henry Kucera vid Browns Universitet, Rhode Island 1963-64. Den innehåller drygt en miljon ord amerikansk engelska.
[4] LOB-korpusen sammanställdes av Stig Johansson och Geoffrey N. Leech vid Oslo Universitet under åren 1970-78. Den innehåller en miljon ord brittisk engelska och motsvarar alltså Brownkorpusen.
[5] Leech (1992:106) påstår att termen *computer corpus linguistics* för första gången dök upp på 80-talet, och anser att den är mer passande för att ange dagens språkstudier som bedrivs med hjälp av datalagrade korpusar, i motsats till den korpuslingvistik som bedrevs redan före datorns tillkomst.
[6] BNC (British National Corpus) t.ex. innehåller drygt 100 miljoner ord.

En korpus användningsområden

Korpusar har bland annat använts flitigt inom lexikografin. En korpus kan nämligen illustrera hur ett ord eller en fras uppträder i ett yttrande, eller illustrera dess syntaktiska egenskaper i en mening.

> Lexicographers […] will use the corpus for information on the actual usage of the words they cover. It may be consulted directly for information on specific words; or it may be processed in various ways, in order to develop parts of a lexical database. (Atkins 1992:13)

Ett stort och känt korpusbaserat ordboksprojekt är COBUILD-ordboksprojektet (Sinclair 1987).[7]

Inom språkforskningen kan en korpus användas i rent beskrivande empiriska studier eller i mer teoretiska studier (se bl.a. Atkins 1992:14). McEnery & Wilson (2001:103-132) nämner en rad olika områden för språkstudier där korpusar kan användas, bl.a. semantiska studier, psykolingvistik, kulturella studier.

Språkvetaren kan använda en korpus som empirisk bas för att illustrera, validera, kontrollera utsagor om språkliga fenomen utifrån en viss teori, eller tvärtom kan han använda korpusen som utgångspunkt för studien och som underlag för en viss teori. Den förra metoden kallas för *corpus-based*, den senare för *corpus-driven* (Ooi 1998:51 som hänvisar till Sinclair). Ooi (ibidem:52) utvidgar Sinclairs uppdelning och kallar det sistnämnda förfaringssättet *top-down* eftersom lingvisten utgår från en teori som han tillämpar på det empiriska korpusmaterialet för att kunna validera eller förkasta den. I den korpusstyrda metoden är arbetssättet *bottom-up*: utifrån empiriska data försöker forskaren komma fram till en teori. Ooi (ibidem:52) tillägger att det i praktiken ofta är nödvändigt att kombinera dessa två arbetssätt.

Är det bara korpusar som gäller?

För- och nackdelarna med korpusanvändning i språklingvistiska studier har flitigt diskuterats.

Till fördelarna hör att korpusar gör det lättare för språkvetaren att komma fram till mer objektiva iakttagelser än de han skulle kunna göra enbart med utgångspunkt i sina egna (meta)språkkunskaper. Intresset för, samt en allt större förlitan på korpusar i språkstudier har delvis vuxit fram ur en insikt om att introspektion inte räcker till.

> The problem about all kinds of introspection is that it does not give evidence about usage. The informant will not be able to distinguish among various kinds of language patterning – psychological associations, semantic groupings, and so on. Actual usage plays a very minor role in one's consciousness of language and one would be recording largely ideas about language rather than facts of it. (Sinclair 1991:39)

Sinclair (1991:5) uttalar sig starkt negativt om språkexempel som är språkvetarens egen uppfinning: "[…] the absurd notion that invented examples can actually represent the language better than real ones."

Att förlita sig på sin subjektiva - underförstått inte nödvändigtvis korrekta - språkintuition innehåller klart en del begränsningar.

[7] Ett exempel på ett korpusbaserat ordboksprojekt som har utarbetats vid Institutionen för Nordiska språk vid Gents Universitet är den nederländsk-danska ordboken (i tryck2002) (Laureys & Rawoens 2001).

> Since the average person's own mental lexicon is firstly, finite and secondly, static and unchanged (i.e. in need of updating) it is probably not sufficient to rely on only one person's linguistic intuitions. (Ooi 1998:48)

Fördelen med en korpus är nämligen att den i själva verket återspeglar ”the collective intuitions of a relevant group of people using the word or linguistic expression under study.” (ibidem 1998:49).

Stubbs (1996:28) är också tydlig på den punkten: ”Languages should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences.”

Vid utforskningen av ett främmande språk där forskaren inte har modersmålskompetens erbjuder en korpus också den mycket stora fördelen att man kan komma åt en mängd autentiskt material på det främmande språket (Leech 1991b:74).

Många (bl.a. Svartvik 1992:10) nyanserar dock den positiva bilden genom att säga att alltför stor förlitan på korpusar kan innebära vissa farhågor.

Summers (1996:266) varnar för risken att blint förlita sig på en korpus, och säger att undersökningen gärna får vara ”corpus-based, but not corpus-bound.” En korpus är inte nödvändigtvis ”felfri”, precis som det verkliga språkbruket som den återspeglar.

Dessutom bör man akta sig för rent kvantifierande studier som bara bekräftar det man redan vet.

> […] counting occurrences, in a large number of cases, is merely a laborious way of coming to conclusions one has already arrived at subjectively. (Leech 1966:73)

Det är inte heller alltid bäst med ”ju större, desto bättre” korpusar (bl.a. Svartvik 1992:10). Leech (1996:10) varnar för att det är naivt ”to focus merely on size”, fast han förnekar inte att det är bekvämt att ha tillgång till stora korpusar (ibidem 1996:13). Hur stor en korpus än är ska man vara medveten om dess begränsningar: en korpus kan aldrig vara till hundra procent representativ för ett helt språk eftersom den alltid bara omfattar en del av språket vars möjligheter i själva verket är obegränsade: ”It is evident that no corpus can represent the language as a whole” (Stubbs 1995:50).

Frågan om storleken hänger naturligtvis också ihop med studieobjektets förekomstfrevens. Ju mindre frekvent en viss språklig företeelse är, desto större korpus behövs. Det är viktigt att få tillräckligt många träffar för att resultatet ska kunna signifikansprövas (Sinclair 1991:18-19) (Atkins et al. 1992:5). Förekommer en viss språklig företeelse inte i förväntad grad behöver detta inte omedelbart leda till slutsatsen att denna företeelse inte är frekvent i hela språket. Church et al. (1991:124) varnar för risken med detta fenomen som han kallar ”failure-to-find fallacy”:

> That is, when you don't have much evidence for something, it is very hard to know whether it is because it doesn't happen, or because you haven't been looking for it in the right way (or in the right place).”

Detta varnar också McEnery & Wilson (2001:30) för: ”more common utterances might be excluded simply by chance”. Denna problemställning anknyter till frågan om en korpus delrepresentativitet för ett helt språk. Denna delrepresentativitet kan man inte ändra på, inte ens genom att sammanställa en mycket stor korpus. Det man däremot kan bearbeta är korpusens interna kvalitet och representativitet. Man kan begränsa sig till en korpus som representerar en viss språkvariation (McEnery & Wilson 2001:30) och klargöra detta i anslutning till undersökningen.

När det gäller storleken finns vissa riktlinjer om hur stor en korpus bör vara för att den ska kunna tjäna som en pålitlig empirisk bas. För vissa studieområden kan en korpus på 1 miljon ord räcka (se t.ex. Sinclair 1991:24). Lauridsen (1996:67) nämner exempelvis kontrastiva studier för LSP (*Languages for Special Purposes*). Hon tillägger dock att en korpus på 1 miljon ord är långt ifrån tillräcklig för lexikografiskt arbete. Detta påpekar också Leech (1991:75) när han talar om Brown Corpus (som innehåller 1 miljon ord):

> While the size of the Brown Corpus may be considered adequate to the study of common features (….), it is manifestly inadequate as a resource for (for example) lexicography, since the corpus contains only c. 50,000 word types, of which c. 50 per cent occur only once in the corpus.

Sinclair (1991:100) håller med om att kravet på en korpus storlek hänger ihop med studieobjektet:

> The received wisdom of corpus linguistics is that fairly small corpora, of one million words or even fewer, are adequate for grammatical purposes, since the frequency of occurrence of so-called grammatical or function words is quite high.

Vill man däremot studera något som t.ex. kollokationer räcker en relativt liten korpus inte till:

> For the study of collocation and phraseology, it is necessary to study huge amounts of text in order to isolate the recurrent patterns and reduce the prominence of the transitory ones. (Sinclair 1991:20)

Överhuvudtaget favoriserar Sinclair starkt stora korpusar, till och med korpusar med obegränsad storlek:

> It is […] necessary to have access to a large corpus because the normal use of language is highly specific, and good representative examples are hard to find. This is as true of grammar as of lexis, because grammar is not made of just the patterns of the common grammatical words, but relies on the whole vocabulary of the language. (Sinclair 1991:100)

Att en korpus har ett begränsat omfång brukar betraktas som ett typiskt kännetecken för korpusar. Däremot finns det också så kallade monitorkorpusar som inte omfattar något begränsat antal ord, utan som hela tiden utökas med nytt språkmaterial och som bara blir större och större. En sådan korpus sammanställdes av Sinclair (1991:18): ”The only guidance I would give is that a corpus should be as large as possible, and should keep on growing.” Monitorkorpusar lämpar sig bra för lexikografiskt arbete (som t.ex. Sinclairs arbete med COBUILD-ordboken). Nackdelen är dock att de inte lämpar sig för kvantitativa studier (McEnery & Wilson 2001:30-31).

Ger en korpusundersökning inte tillräckligt många träffar eller visar träffarna sig vara kvalitativt otillräckliga, kan lingvisten bestämma sig för att supplera korpusundersökningen med introspektion och/eller elicitering[8]. I princip kan språkforskaren hantera tre olika sorters infallsvinklar: han kan förlita sig på sin egen språkkompetens (introspektion), han kan vända sig till informanter och undersöka deras passiva och aktiva språkbruk (elicitering)[9] eller han kan använda sig av insamlat språkmaterial (korpusanalys) (Lauridsen 1989:118). Svartvik (1992:8) menar att det är bra att diversifiera och att ”all these methods should be seen as complementary.” För Leech (1991:74) står det klart att korpuslingvistik i själva verket är ”a question of corpus

[8] Om supplering genom elicitation se också De Mönnink (1997:227).
[9] En mer informell och mer förekommande variant är där språkforskaren vänder sig till en informant, t.ex. modersmålstalare, som han för en mer spontan diskussion med (Sinclair 1991:39).

plus intuition, rather than a corpus *or* intuition." Att de tre ovannämnda metoderna inte behöver utesluta varandra menar också Ooi (1998:52):

> It is important to note that a reliance on corpus data does not mean e denial of the other two methods for the gathering of lexicographic or lexical evidence. Far from exclusiveness, it is often necessary to utilize all three methods for the gleaning of such evidence adequately.

En kritisk syn och subjektiv reflektion på korpusbaserade resultat är alltid förnuftig: "Your intuition about language […] is a most important asset." (Sinclair 1997:32)

Kanske viktigare än att ha en stor korpus är att ha en diversifierad korpus, där materialet har hämtats från många olika talare eller källor, och där texterna har en hög representativitetsgrad för det språk de ämnar återspegla. Detta anknyter till representativitetskriteriet för en korpus (McEnery & Wilson 2001:29): "In linguistics, we are often more interested in a whole variety of a language, rather than in an individual text or author." Sinclair (1991:18) säger att "The diversity of sources is an essential safeguard".

Lauridsen (1989:199) beskriver "et eksemplarisk korpus" som följande:

a) et korpus baseret på en plausibelt lydende, hypotetisk fordeling af bestemte tekster i den objektive, altså den virkelige verden
b) et korpus baseret på en veldefineret, eller, som det reelt vil blive, en approximativt (*sic!*) defineret delmængde af et sprog (sml. Bergenholz 1988:231)

Många korpusar är dock inte direkt diversifierade: det språkmaterial de innehåller har ofta hämtats från likartade källor eller genrer. Ett exempel är korpusar som består av endast pressmaterial eller endast skönlitterära texter. I sådana korpusar kan språket vara präglat av en skribents eller författares skrivstil. Detta varnar bl.a. Sinclair (1991:17) för: "If we are to approach a realistic view of the way in which language is used, we must record the usage of the mass of ordinary writers, and not the stray genius or the astute journalist." Detta bör språkforskaren vara medveten om och vara försiktig med om han vill göra mer allmänt gällande uttalanden om hur ett språk beter sig.

Har lingvisten inte tillgång till en diversifierad korpus kan en lösning dock vara att matcha de undersökningsresultaten från en bestämd korpusundersökning mot en annan korpus eller flera andra korpusar (Stubbs 1995:50). Också Woods (1986) understryker att det är viktigt att hypoteser provas på flera olika sätt.

Sammanfattningsvis kan sägas att språkforskaren bör eftersträva att ha en korpus som är "so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study." (Atkins et al. 1992:6). En korpus kan ta många olika former, att välja "rätt" korpus hänger till stor del ihop med den sortens undersökning man vill bedriva. Till slut är det viktigt att vara medveten om att de korpusbaserade undersökningsresultaten relaterar till korpusens natur.

Typologi av korpusar i monolingual och kontrastiv språkforskning

Enspråkiga korpusar lämpar sig för olika former av monolingual språkforskning. Vid diakronisk språkforskning bör en korpus innehålla textmaterial från olika tidsperioder[10]. Vid stilistisk språkforskning ska korpusen helst innehålla texter från specifikt utvalda författare eller från vitt skilda textgenrer.

Är man intresserad av talspråkliga drag måste man förstås anlita en talkorpus. Korpusar över talat språk är relativt nya. En talkorpus innehåller vanligtvis både ljud och text (med andra ord transkriptionen av det som sägs). Exempel på korpusar över talat språk är CGN (Corpus Gesproken Nederlands)[11], och Swedish Spoken Language Corpus[12].

För kontrastiv språkforskning kan man använda sig av flerspråkskorpusar, som består av textmaterial på två eller fler språk.

Kontrastiv språkforskning innebär att man studerar två eller flera språk genom att kontrastera dem mot varandra. Avsikten med en kontrastiv språkstudie är inte bara att belysa specifika skillnader mellan dessa språk, utan också att belysa språkens respektive individuella egenskaper (se bl.a. Johansson 2000:4 ; Aarts 1998:Introduction ; Aijmer & Altenberg 1996:12).

> The confrontation of languages is important from the point of view of translation theory, language typology and the study of language universals. Above all, it can be an excellent way of highlighting the structure of the languages compared. This means that CA [contrastive analysis] could be an aid in formulating accurate descriptions of individual languages. (Johansson 1975:15)

Kontrastiva språkanalyser kan tillämpas på en rad olika områden som språkundervisning, tvärspråklig analys och lexikografiskt arbete (Lauridsen 1996:64).

Det finns olika sorters flerspråkiga korpusar. Det råder en viss förvirring angående benämningen på de olika korpusarna. Här har jag valt att använda den terminologi som verkar vara mest transparent och som andra (bl.a. Johansson 1998, Lauridsen 1996) inom det nordiska språkområdet följt.[13] Förutom så kallade översättningskorpusar, som innehåller originaltexter på ett visst språk plus deras översättningar till ett annat språk, finns det så kallade parallellkorpusar som inte innehåller översättningar, utan texter skrivna på olika originalspråk som är jämförbara på så sätt att de hämtats från en viss genre eller ett visst specialområde. [14]

Hur bestämmer man sig för vilken slags flerspråkig korpus man ska använda? Johansson (2000:4) hänvisar till och håller med Carl James som säger att ”translation is the best basis of comparison”. Han fortsätter med att säga att ”The use of multilingual corpora, with a variety of texts and a range of translators represented, increases the validity and reliability of the comparison.”

Men andra uppfattningar finns likaså. Lauridsen (1996:65) undrar: ”Can one really compare language structure or text structure on the basis of translations?”

Aarts (1998:Introduction) säger att:

---

[10] Vid diakronisk språkforskning har man i övrigt knappt något annat val än att använda sig av en korpus.
[11] För en projektbeskrivning se http://lands.let.kun.nl/cgn/home.htm
[12] För en projektbeskrivning se http://www.ling.gu.se/projekt/SLSA/SLcorpus.html
[13] Se också Johansson (1998:4-5 fotnot) för denna diskussion.
[14] Detta är alltså i motsats till t.ex. McEnery & Wilsons (2001:70) definitioner som kallar ”parallel corpora” (parallellkorpusar) för korpusar som innehåller originaltexter på ett visst språk plus deras översättningar till ett annat språk. ”Translation corpora” eller ”comparable corpora” innehåller enligt deras definition inga översättningar, utan texter skrivna på olika originalspråk som är jämförbara i stil och genre. Gellerstam (1996:54) använder sig även av termen ”parallel texts” antingen för att ange översättningstexter eller texter som är jämförbara i stil och genre.

> Full comparability can only be achieved in translation corpora, i.e. corpora containing texts from a source language and their translations into one or more target languages.

Han nyanserar dock sin uppfattning:

> [This] does not, unfortunately, imply that it is also a perfect research tool for linguists who want to compare two or more languages. An intrusive factor in such corpora is the translation activity itself, which may affect the texts of the target language.

Är man specifikt intresserad av att studera översättningsprocessen är översättningskorpusar klart det mest lämpliga. Man kan även välja att undersöka just påverkan av källspråket på målspråket. Gellerstam (1996:53-54) kallar dessa spår av tydlig påverkan för *translationese*.

Korpusar i ett svensk-nederländskt kontrastivt forskningsprojekt

Efter ovanstående allmänna överväganden och hänvisningar till litteraturen vill jag nu gärna ange några personliga skäl till att använda korpusar inom ramen för mitt avhandlingsarbete, och ge en kortfattad beskrivning av de olika korpusar som jag använder mig av.

Det främsta skälet till att jag i min undersökning utgår från korpusar är mitt intresse för att studera verkligt språkbruk. Jag kommer dock att komplettera korpusanalysen med eliciteringstest som kan bidra med ytterligare upplysningar angående språkattityder.

Till att börja med är det viktigt att etablera ett tertium comparationis som utgör ett icke-språkspecifikt underlag för analysen. I min avhandling utgör kausativitet utgångspunkten för en kontrastiv analys i nutida svenska och nederländska, där jag undersöker hur detta (semantiska) begrepp kan uttryckas med hjälp av grammatiska kategorier, i synnerhet analytiska kausativa verbkonstruktioner. Jag utgår från en typologi över sådana konstruktioner i nederländskan respektive svenskan med tonvikt på deras egenskaper på syntaktisk, semantisk och pragmatisk nivå. Denna typologi prövar jag ut mot nederländska och svenska monolinguala korpusar. Att jag har valt att tillämpa en korpusbaserad snarare än korpusstyrd metod hänger dels ihop med att analytiska kausativa verbkonstruktioner inte förväntas ha någon hög förekomstfrekvens i de olika korpusarna, som visserligen är diversifierade, men som ändå återspeglar ett ganska ”allmänt” språk[15]. Detta behöver dock inte innebära att korpusanalysen inte kan lyfta fram kanske oväntade mönster som kan styra den teoretiska modellen. Detta kan med andra ord betyda att korpusanalysen också kan få inslag av en mer korpusstyrd metod.

De monolinguala korpusarna erbjuder möjligheten att studera stora mängder naturligt språkmaterial (språkbruk) och att kartlägga kausativkonstruktionerna i deras kontext, granska hur de uppträder och hur deras spridning förhåller sig till textstrukturen. När det gäller de monolinguala korpusarna har jag valt att använda mig av redan befintliga korpusar. Dessa är i första hand korpusar med skriftligt textmaterial[16], närmare bestämt Språkbankens korpusar[17] för svenskans del, och INL-subkorpusarna[18] för nederländskans del.

[15] Jag använder mig alltså inte av korpusar som återspeglar ett mera specifikt subspråk som kan förväntas innehålla fler kausativa verb och verbkonstruktioner som t.ex. ett tekniskt eller vetenskapligt språk, där kausala eller resultatsbundna förhållanden kanske oftare beskrivs.
[16] Förutom dessa kommer jag också att använda mig av de monolinguala talkorpusarna för svenska (Swedish Spoken Language Corpus) respektive nederländska (CGN), men detta går jag inte närmare in på här.
[17] Se http://spraakbanken.gu.se/
[18] Se http://www.inl.nl/

INL-korpusarna innehåller tre subkorpusar med endast icke-litterärt material. Förutom en komponent med juridiskt textmaterial innehåller de övervägande pressmaterial från 90-talet. Jag använder mig av subkorpusarna på 27 miljoner ord respektive 38 miljoner ord, sammanlagt 65 miljoner ord. Den förstnämnda subkorpusen innehåller texter från en nederländsk tidning (NRC Handelsblad), den senare subkorpusen är mer diversifierad och innehåller förutom tidningstexter från en tidning (Meppeler Courant) också en komponent med juridiskt material, plus en blandgrupp där också tidningstexter från en belgisk (flamländsk) tidning ingår (De Standaard).

Språkbankens korpusar innehåller totalt drygt 90 miljoner ord. Man kan särskilja flera subkorpusar med både litterärt och icke-litterärt textmaterial.

De litterära subkorpusarna utgör 20 procent av den totala korpusen och innehåller sammanlagt 18 miljoner ord. Bland de litterära texterna kan man särskilja en delkorpus med äldre litterära texter på sammanlagt drygt 8 miljoner ord (bl.a. romaner av C.J.L. Almqvist, H. Bergman, romaner och brev av A. Strindberg) och en delkorpus med modern skönlitteratur på drygt 9 miljoner ord (Bonniers romaner med svenska som originalspråk bl.a. A. Lundkvist, K. Ekman samt översatt litteratur på svenska bl.a. A. Christie, M. Kundera).

Subkorpusarna med icke-litterärt textmaterial innehåller sammanlagt drygt 74 miljoner ord och utgör 80 procent av hela Språkbankens korpus. Till de icke-litterära subkorpusarna hör subkorpusar med pressmaterial (sammanlagt 40 miljoner ord), riksdagsprotokoll (4 miljoner ord), lagtexter, lexikaliskt material från Svenska Akademiens ordbok (23,6 miljoner ord), Svenska Akademiens ordlista (400 000 löpord) och några mindre subkorpusar på ett par tiotusen ord.

Jag gjorde ett urval ur alla subkorpusar och selekterade bort ett antal subkorpusar (det skönlitterära materialet och de subkorpusar som innehåller icke-litterärt textmaterial annat än pressmaterial) så att jag till slut fick ihop en subkorpus som skulle vara jämförbar med INL-korpusen i både omfång och textgenre. Detta resulterade i en subkorpus som endast innehåller pressmaterial med tidningstexter från 60-talet fram till 90-talet, fast den övervägande delen (84%) är från svenska tidningar från 90-talet (Dagens Nyheter, Göteborgs-Posten, Svenska Dagbladet), sammanlagt 40 miljoner ord.

För den specifikt kontrastiva infallsvinkeln i undersökningen bestämde jag mig för att i samråd med Språkbanken bygga upp en bilingual svensk-nederländsk korpus bestående av en litterär och en icke-litterär subkorpus med både originaltexter och översättningstexter. En bilingual korpus lämpar sig för kartläggning av hur kausativkonstruktionerna har översatts åt båda hållen, och för granskning av förhållandet mellan de olika kausativkonstruktionerna och de kausativa verben. En kontrastiv undersökning med utgångspunkt i den bilinguala korpusen kan också ses som en förlängning av den monolinguala korpusanalysen. Dels kan den ev. bidra med extra upplysningar som inte kommer fram vid den monolinguala korpusanalysen, dels kan subkorpusen med originaltexterna användas som utökning av de monolinguala korpusarna och bidra med en mer diversifierad infallsvinkel (skönlitteratur visavi pressmaterial).

SALT-projektet

Den svensk-nederländska korpusen håller jag på att sammanställa i samarbete med doktorander från Göteborgs universitet inom ramen för projektet SALT (Språkbankens arkiv för länkade texter)[19]. Projektet samordnas av

[19] För SALT-projektbeskrivningen se: http://spraakbanken.gu.se/lb/salt/

Språkbanken[20] och målet är att bygga upp ett antal parallellkorpusar med svenska som ett av språken och med ett annat främmande språk som det andra språket i varje respektive delkorpus.

***SALT** Språkbankens arkiv för länkade texter*		
	engelska	
tyska		ryska
	svenska	
nederländska		franska
	italienska	

Figur 1: Strukturen i SALT-korpusarna

De olika delkorpusarna innehåller skönlitterärt textmaterial[21] och originaltexter på svenska plus deras översättningar till det andra språket (t.ex. ryska eller nederländska), samt originaltexter på det främmande språket och deras översättningar till svenska.[22] De här parallellkorpusarna är på så sätt egentligen en blandning av parallell- och översättningskorpusar, men kallas inom ramen för SALT för parallellkorpusar. Den bidirektionella strukturen i SALT-korpusarna skapar möjlighet till en rad olika undersökningar. Det är inte bara möjligt att jämföra originalspråkstexter och deras översättningar, utan man kan också jämföra de båda originalspråken och till och med originaltexterna på ett visst språk med översättningstexterna till samma språk.[23]

svensk-nederländsk parallellkorpus	
svenska original	nederländska original
svenska översättningar	nederländska översättningar

Figur 2: Strukturen i den svensk-nederländska korpusen

SALT-korpusarna innehåller hela verk i stället för mindre textavsnitt på ett bestämt antal ord[24]. Att språket i korpusen som helhet härmed inte är lika diversifierat som i en korpus med olika mindre utdrag ur texten kan betraktas som en nackdel: korpusarna representerar ett fåtal författar- och översättarstilar.

> The penalties to pay for including whole documents are that in the early stages of gathering, the coverage will not be as good as a collection of small samples and the peculiarities of an individual style or topic may occasionally show through into the generalities. (Sinclair 1991:19)

Till fördelarna hör att det blir möjligt att använda delar av hela korpusen för mindre undersökningar där en enda roman kan användas som undersökningsmaterial. Att språket har en enhetlig form och stil romanen igenom kan också ses som en fördel med en korpus bestående av hela texter.

[20] Projektet finansieras dels av Humanistiska Fakulteten vid Göteborgs universitet, dels av Riksbankens Jubileumsfond.
[21] Den svensk-nederländska delen kommer även att innehålla icke-litterärt material.
[22] Strukturen i SALT-korpusarna är jämförbar med den i ESPC http://www.englund.lu.se/research/corpus/corpus/espc.html (Aijmer et al 1999:79-80) och ENPC (Johansson http://www.hit.uib.no/hit/enpc.htm).
[23] Se också Aijmer (1996:82) om ESPC.
[24] Detta till skillnad från t.ex. ESPC och ENPC som innehåller textavsnitt på ca 15,000 ord (se respektive projektbeskrivning jfr fotnot 20).

> Not many features of a book-length text are diffused evenly throughout, and a corpus made up of whole documents is open to a wider range of linguistic studies than a collection of short samples. (Sinclair 1991:19)

Den svensk-nederländska korpusen kommer att vara den första i sitt slag. Den sammanställs enligt gemensamma SALT-principer[25] och kommer att innehålla både en litterär och en icke-litterär subkorpus på sammanlagt ca 3 miljoner ord.[26]

Uppbyggnad av den svensk-nederländska parallellkorpusen

Här nedan kommer jag att belysa det konkreta arbetet med att bygga upp denna parllellkorpus. Arbetet sker i enlighet med Språkbankens riktlinjer och enligt vissa arbetsmallar (t.ex. för uppmärkning av materialet).[27]

Valet av texter

Valet av korpustexterna baseras på Språkbankens riktlinje om att korpusen ska avspegla olika typer av texter, den ska närmare bestämt innehålla skönlitteratur och facklitteratur från senare hälften av 1900-talet. Litterära texter, och närmare bestämt romaner, innehåller vanligtvis en blandning av prosatext och dialoger. Språket är oftast varierat och återspeglar ett ganska vanligt levande språk från en viss tidsepok. En icke-litterär subkorpus utgör en motvikt med ett annat slags språk. Detta ska dock helst inte representera en enda sorts specialiserat språk, utan vara diversifierat i och med att olika specialområden finns representerade.

En första begränsning som vi fick räkna med när vi försökte ta fram en lista över vilka verk vi ville ha med i korpusen, var att vi bara kunde utgå från befintliga nederländska översättningar av svenska verk samt svenska översättningar av nederländska verk. Till den litterära subkorpusen selekterade vi, i enlighet med Språkbankens riktlinjer, ett antal verk från båda språkområdena, som återspeglar modern litteratur (inte tidigare än 1960-talet).[28] Det blev sammanlagt 26 romaner, varav 13 originaltexter (7 svenska original, 6 nederländska original) och 13 översättningar. Bland de romaner som finns med i korpusen finns t.ex.: Bergman, I., ”*Laterna Magica*”, Claus, H., ”*De Geruchten*”, Krabbé, T., ”*Het Gouden Ei*”. Den litterära subkorpusen kommer att innehålla 2 miljoner ord.

Till den icke-litterära subkorpusen valdes ett antal verk från olika specialområden så att språket skulle vara diversifierat: psykologi, vetenskapshistoria, naturvetenskap, geografi, socialhistoria. Några av verken som finns med i den icke-litterära subkorpusen är Cullberg, J., ”*Dynamisk psykiatri i teori och praxis*”, Dekker, R., van de Pol, L., ”*Vrouwen in mannenkleren*”. Den icke-litterära subkorpusen kommer att innehålla 12 verk (4 svenska original, 2 nederländska original), sammanlagt 1 miljon ord.[29]

[25] Detta är både innehållsmässiga (bl.a. vilka texter som ska ingå i korpusen) och formella (berör bl.a. uppmärkning) principer.
[26] Språkbanken föreskriver: ”Ett lämpligt riktmärke när det gäller uppbyggnad av korpus är en miljon löpande ord på svenska och lika mycket på det främmande språket.” (projektansökan till RJ). Vi valde att ha en lite större korpus som skulle kunna lämpa sig för olika ändamål, också för studier av mindre frekventa språkföreteelser som kausativa verb. Den interna fördelningen är följande: 2 miljoner ord litterära texter, 1 miljon ord icke-litterära texter.
[27] Arbetet med den svensk-nederländska korpusen startade i januari 2001 och beräknas vara klart i december 2002.
[28] SALT-projektet har som mål att ha tre svenska originalverk som finns i översättning till samtliga ingående särspråk.
[29] För samtliga verk gäller att de kommer att ingå i den svensk-nederländska korpusen under förutsättning att upphovsrättsfrågan är lost.

Skanning

Förutsättningen för att man ska kunna använda en datalagrad korpus är naturligtvis att den finns tillgänglig i elektronisk och sökbar form. Med tanke på dagens textmängder som kommer i elektroniskt format, kan man lätt förvånas över att det innebär mycket mer möda än vad man skulle tro för att komma åt materialet. Moderna litterära texter ligger inte bara ute på nätet. Skälet till detta är bl.a. att de skyddas av upphovsrättslagen, som dessutom innebär att man inte bara fritt får kopiera textmaterial[30]. Vill man få romanerna i elektronisk form, får man helt enkelt skanna dem[31], förutsatt att man av förlagen har fått tillstånd att använda materialet inom ramen för forskningen. Resultatet av skannandet är en *.tif*-fil – som är ett vanligt format för bilder (jfr *.gif*-filer).

Efter själva skannandet är det nödvändigt att köra ett så kallat OCR (Optical Character Recognition)-program (t.ex. Omnipage Pro 10) som tolkar bild till text. Detta är nödvändigt för att man senare ska kunna arbeta med och bearbeta texten. Detta innebär att *.tif*-filen laddas in i programmet som förvandlar den till en *.opd*-fil[32], vilket är ett slags textformat.

Korrekturläsning

Har man kommit så långt, börjar den del av arbetet som är mest mödosam: korrekturläsningen. Programmet som tolkat bild till text kommer nämligen ofta att ha misstolkat.[33] Det enklaste är att man sparar texten i *.rtf*-format för att kunna fortsätta korrigera i ett ordbehandlingsprogram som MsWord.[34] Exempel på vanliga misstolkningar är att en liten prick eller fläck på sidan tolkas som en accent eller en prick ovanpå en viss bokstav: t.ex. ordet *f.ar* (med en liten fläck på sidan) kan misstolkas som *fàr* (med accent). Vissa bokstavskombinationer kan misstolkas: *rn* misstolkas ofta som *m* (vilket t.ex. kan göra att ordet *barn* misstolkas som *bam*), *li* misstolkas lätt som *h*. Avstavade ord där andra ledet fortsätter på nästa sida ska skrivs ihop igen, det vill säga att andra ledet flyttas över till föregående sida. Att korrekturläsa är en mycket viktig fas i arbetet, meningen är ju att texten ska bli användbar och sökbar (vill man senare t.ex. söka på ett ord som *barn* kommer sökprogrammet inte att visa ord som *bam*.)

Uppmärkning av materialet

> In the case of sources which need to be converted from printed form, mark-up will have to be introduced. (Atkins 1992:9)

När texten är korrekturläst är den klar för att märkas upp eller taggas[35]. Uppmärkning eller annotering av en text betyder att man infogar extra information i texten.

> Mark-up […] means introducing into the text, by means of some conventional set of readable labels or tags, indicators of such text features as, for example, chapter, paragraph, and sentence boundaries, headings and titles, various types of

[30] Språkbanken tar på sig arbetet med de upphovsrättsliga frågorna (om upphovsrättsfrågan se också Atkins et al.1992:4 och Sinclair 1991:15).
[31] "For the mass of books printed by conventional methods, scanning is much the best alternative." (Sinclair 1991:14)
[32] Omnipage-format.
[33] Det går att genomföra en del korrektioner redan i Omnipage, så snart programmet upptäcker något konstigt, vanligtvis namn eller alltför långa ord, kommer det en liten ruta där man har möjlighet att rätta. Denna funktion kan jämföras lite grann med stavningskontroll i ett ordbehandlingsprogram, men detta är inte tillräckligt för att få fram en felfri text.
[34] Där kan man då börja med att plocka bort de största eller vanligaste felen i texten med hjälp av Words-sökvillkor (hitta och ersätta) och genom att köra automatisk stavningskontroll.
[35] I princip kan texten taggas i vilken arbetsfas som helst, t.ex. under OCR-fasen, men inom SALT-projektet har man valt att låta taggningen ske i en "post-editing" fas.

> hyphenations, printers' marks, hesitations, utterance boundaries, etc. (Atkins et al. 1992:9)

Detta gör man genom att infoga såkallade taggar eller begränsare.[36] Texten sparas i ASCII-format, som är en enhetlig standard med ett begränsat alfabet (antal tecken) som kan läsas på alla sorters datorer och av alla programvaror[37]. Det finns olika sorters uppmärkning. Den enklaste sorten är att man infogar taggar som anger ordgräns, meningsgräns eller styckegräns. Uppmärkningsspråket som man då använder är vanligtvis SGML (Standard Generalized Markup Language), som man känner igen som taggar i form av hakparenteser: för varje stycke man märker upp, infogar man en starttagg och en sluttagg.

Den uppmärkning som görs i den svensk-nederländska korpusen är i själva verket en förenklad uppmärkning som sedan ändras till en uppmärkning som följer xml-versionen av CES (Corpus Encoding Standard)[38].

De taggar som infogas är bland andra kapitelbegränsare. På ett kapitel som är numrerat eller namngett skriver man t.ex. <chapter><title>JULIA</title></chapter>. Sidnummer anges med hakparenteser, bokstaven *p* och siffran som anger sidnumret t.ex. <p15>.

Sedan finns det en del typografiska särdrag som taggas, bland andra kursiv stil, fetstil, ord/fraser med versaler[39] t.ex.: han menade <it>det</it>.

Ytterligare en sak som taggas är de ställen där det uppträder fel i originaltexten: då kan jag själv gå in och infoga en tagg och en kod (t.ex. förkortning av mitt namn) som anger att det är en personlig korrigering: t.ex. <corr sic="bolckflöjt" resp="gr">blockflöjt</corr>
Alla dessa taggar förs in manuellt.[40]

<chapter><title><it>2</it></title>

<ca>HET</ca> water van de inham was blauw als op een kindertekening. Rex Hofman stond op een grote steen aan de zijkant en keek door het sleutelgat van de rotswanden naar de horizon. Hij vroeg zich af hoe de aarde en de maan het samen klaarspeelden elders in deze zee golven van een meter hoog te veroorzaken terwijl hier deze rimpelloosheid heerste.

Figur 3:Exempel på ett taggat textstycke (KrabbéN1[41], s.26)

Korpusmaterialet kommer i nuläget inte att förses med ordklass- eller syntaktisk annotering. Materialet kommer dock att vara tillräckligt förarbetat så att olika parsers kan tillämpas på det i ett senare skede.

Länkningen

Länkning innebär att originalversionen och översättningen av romanen i fråga länkas ihop, mening för mening. Syftet med detta är att det tvåspråkiga materialet ska bli sökbart på ett visst sätt (se nedan). Länkningsarbetet

[36] När det gäller taggning, och i viss mån korrekturläsning, arbetar jag enligt en mall som Språkbanken har tagit fram.
[37] I detta format kommer uppmärkningen att ske, och texten kommer också att lämnas in till Språkbanken som *.txt*-fil. Taggningen sker i en så kallad text editor som t.ex. Notepad.
[38] Detta kallas också för XCES. Se http://www.cs.vassar.edu/XCES/ och http://www.cs.vassar.edu/CES/ för mer information om XCES och CES.
[39] Detta måste alltså göras eftersom all form för uppmärkning som man känner till från t.ex, Word (typsnitt, kursiv) går ”förlorad” i ASCII-format.
[40] De taggar som inte ska infogas är styckebegränsare, eftersom de kan sättas in automatiskt av en datalingvist vid Språkbanken. Jag måste dock ange de ställen där det ska infogas styckebegränsare, med hjälp av radbrytning och två tabbar.
[41] Exemplet har hämtats ur en roman av Krabbé. Koden N1 anger att originalspråket (1) är nederländska (N).

sköts av en datalingvist vid Språkbanken och sker automatiskt med hjälp av ett länkningsprogram (Vanilla Aligner, Unix-baserat[42]) som har utvecklats vid Institutionen för svenska språket av Danielsson och Ridings.[43] Före länkningsprogrammet körs har dock ett meningssegmenteringsprogram körts som fungerar på så sätt att det försöker hitta meningsgränser. Det är bl.a. därför som det är så viktigt att ha genomfört en bra korrekturläsning. Har man t.ex. förbisett ett ställe där en punkt av OCR-programmet uppfattas som kommatecken, hittar meningssegmenteringsprogrammet inte slutet på en mening, vilket kan göra så att länkningsprogrammet kan gå fel.

När ett textpar har länkats är det nödvändigt att gå igenom några manuella korrigeringar som gjorts under själva länkningsarbetet. Förutom detta ska också en lista över möjliga fellänkningar kontrolleras, dvs. icke-1-1-länkar (en mening L1 - en mening L2), som t.ex. 1-2 (dvs. fall där en mening länkats till två meningar: det kan stämma ifall översättaren helt enkelt valt att översätta en mening med två enskilda meningar) eller också 1-0 (dvs. fall där motsvarande översättning verkar saknas, vilket trots allt förekommer överraskande ofta).

Korpusanvändning

Sökning

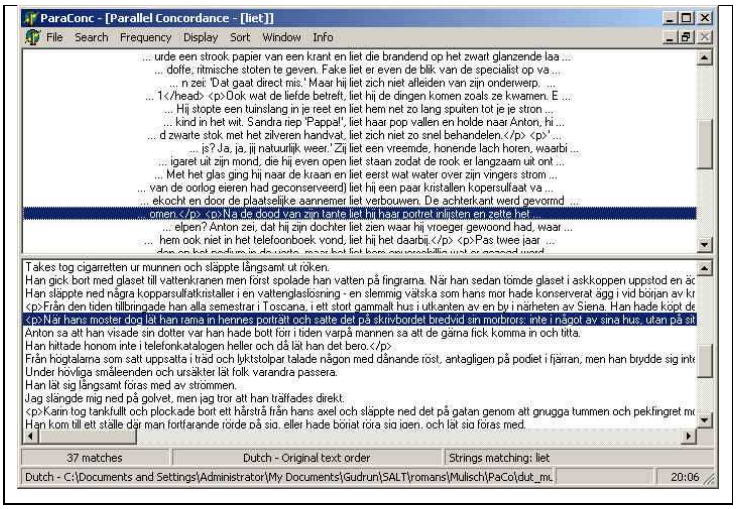
Har man valt att använda en textkorpus i elektronisk form, vill man kunna söka efter språkliga enheter, och helst få fram konkordanser. Ett typiskt presentationssätt är där man får se sökningen ”highlighted” med dess kontext både till vänster och till höger, och olika konkordansrader där ordet man sökt på förekommer. För att få fram konkordanser på så sätt behövs naturligtvis program.

Vissa program för konkordanssökningar installeras lokalt, man laddar in en korpus som man sedan kan söka i. Andra korpusar har gjorts sökbara genom Internet, som BNC (British National Corpus)[44] för engelskan och Språkbankens korpusar för svenskan[45].

Användning inom ramen för forskningen: en svensk-nederländsk kontrastiv studie av kausativa verb

Det konkordansprogram som för närvarande används inom SALT-projektet heter ParaConc[46] och är avsett för parallellställda texter. Programmet fungerar på så sätt att de parallellställda texterna laddas in i programmet och visas samtidigt på skärmen. Söker man någonting på det ena språket, får man se motsvarande översättningar på det andra språket.

[42] Se bl.a. http://spraakdata.gu.se/lb/tools.html
[43] Länkningsprogrammet som är skrivet av Danielsson och Ridings är i själva verket en implementering av den av Church & Gale (1993) föreslagna länkningsalgoritmen.
[44] Se http://thetis.bl.uk
[45] Se http://spraakbanken.gu.se/lb/konk/
[46] Programmet togs fram av Michael Barlow. En beskrivning finns på http://www.ruf.rice.edu/~barlow/parac.html Utveckling av ett söksystem för SALT-parallellkorpusarna pågår vid Språkbanken.



Figur 4: ParaConc skärmbild

I min undersökning om kausativa verb fokuserar jag i första hand inte på de lexikala kausativa verben som t.ex. *orsaka* eller *sätta*, utan på de kausativa verb som utgör kärnan i de analytiska kausativa konstruktionerna som har strukturen: kausativt verb + infinitivkomplement.

Innan jag sätter igång är det viktigt att ta fram en rad paradigm för att undvika att undersökningen begränsas till en ren frekvenslista. Det är alltså en fördel att analysera undersökningsobjektet i källspråket och definiera relevanta semantiska och syntaktiska funktioner som kan utgöra utgångspunkt för att differentiera ytterligare.

Såväl på nederländska som på svenska finns fler kausativa verb som kan utgöra kärnan i en perifrastisk kausativ konstruktion med infinitivkomplement. En målsättning i avhandlingsarbetet är att undersöka vari skillnaderna ligger. En förundersökning (Rawoens 2002) har visat att fördelningen till stor del hänger ihop med subjektets semantiska roll: subjektsreferenten kan fungera som agens eller orsak i högre eller lägre grad och är i enlighet härmed oftast mänsklig respektive icke-mänsklig.

På nederländska utgör verben *doen* och *laten* kärnan i de perifrastiska kausativa konstruktionerna. De följs av infinitivkomplement utan infinitivmärke. Jag illustrerar användningen av dessa två verb med exempelmeningar från den svensk-nederländska korpusen.

- *doen* + NP + V_{inf} : […] de plaag die in haar dorp de mensen sneller, vreemder, onverwachtser *doet* sterven […].(ClausN1)
- *laten* + NP + V_{inf} : Bij 8-7 voor Frankrijk *liet* ze hen van speelhelft wisselen. (KrabbéN1)

Deras direkta motsvarigheter på svenska är de kausativa verben *få* och *låta*, men även verben *komma, ha* och *förmå* kan uppträda som kärna och som alternativ till *få.* I konstruktionerna med *få, komma, ha* och *förmå* tar resultatspredikatet infinitivmärket *att*. Dessa är de möjliga mönstren på svenska:

- *få* + NP + *(till)* + *att* + V_{inf} : Jag visste […] att jag *fick* människor att göra som jag ville, […].(BergmanZ1)
- *låta* + NP + V_{inf} : Han köpte ett kilo äpplen och *lät* expediten stoppa dem i en plastkasse.(KrabbéZ2)
- *komma* + NP + *(till)* + *att* + V_{inf} : Där fanns särskilt en bild som *kom* mig att yla av sorg.(BergmanZ1)
- *ha* + NP + *(till)* + *att* + V_{inf} : Hon *hade* honom att bygga ett nytt garage. SAG

- *förmå* + NP + *(till)* + *att* + V_{inf} : Jag *förmådde* honom att söka psykiatrisk vård - ingenting hjälpte.(BergmanZ1)

En viktig semantisk skillnad mellan de nederländska verben *doen* och *laten* är att *doen* används för att uttrycka relativt enkla, snarare direkta kausala förhållanden. *Laten* däremot uttrycker att den kausala relationen snarare är indirekt (Verhagen 1997:70). Subjektsreferenten till *doen* kan antingen vara mänsklig eller icke-mänsklig, till *laten* är den vanligtvis mänsklig.

I vissa fall är skillnaderna mellan dessa två verb mycket tydliga, i andra fall är gränsen snarare vag och i några enstaka fall kan verben t.o.m. vara utbytbara.

Subjektsreferenten till *få* är antingen mänsklig eller icke-mänsklig. Subjektsreferenten till *låta* är vanligen mänsklig. *Komma* tar icke-mänskliga subjektsreferenter. Subjektsreferenterna till *ha* och *förmå* är vanligen mänskliga.

Skillnaderna mellan de svenska kausativa verben är av ett annat slag än skillnaderna mellan de nederländska kausativa verben. Verben *få* och *låta* är aldrig överlappande på samma sätt som *doen* och *laten* är. Däremot kan verben *komma, ha* och *förmå* förekomma som ”alternativ” till verbet *få*. Dessa tre verb förekommer dock inte i lika stor utsträckning som *få* och deras semantiska begränsningar är större än när det gäller *få*: de kan uttrycka olika grader av påverkan och modalitet (Altenberg 2001:102) och är ofta mer formella.

Har man kartlagt dessa verb på båda språken, kan man börja titta på hur de översätts från det ena till det andra språket och på vilka sätt deras egenskaper överförs genom översättningsprocessen.

Tar man det nederländska kausativa verbet *laten* som utgångspunkt, ser man att de motsvarande svenska översättningarna ofta innehåller verbet *låta*, som i de här korpusexemplen:

Hij kocht een kilo appels en *liet* die in een plastic tasje doen. KrabbéN1

Han köpte ett kilo äpplen och *lät* expediten stoppa dem i en plastkasse. KrabbéZ2

Bij 8-7 voor Frankrijk *liet* ze hen van speelhelft wisselen. KrabbéN1

Vid 8–7 för Frankrike *lät* hon dem byta planhalva. KrabbéZ2

Det nederländska kausativa verbet *doen* har ibland översatts till svenska med en konstruktion med det kausativa *få,* men väldigt ofta, har det översatts med ett lexikalt verb.

Of, om de buren niet te *doen* schrikken, […] ClausN1

Eller, för att slippa *skrämma upp* grannarna, […] ClausZ2

[…] de dreun *deed* de grond beven. BergmanN2

[…] dånet *skakade* marken. BergmanZ1

Till slut kan man titta på hur de svenska kausativa verben som t.ex. *få* och *låta* återges i nederländsk översättning. Då får vi fram att den svenska konstruktionen med *få* ofta översätts med en konstruktion med *doen*, men ibland också med *laten* som i det här exemplet.

Jag visste att jag ägde övertalningsförmåga, att jag *fick* mänskor att göra som jag ville, […]. BergmanZ1

Ik wist dat ik over overredingskracht beschikte, dat ik mensen kon *laten* doen wat ik wilde, […]. BergmanN2

När de övriga svenska kausativa verben översätts till nederländska blir resultatet ofta en konstruktion med *doen*.

Där fanns särskilt en bild som *kom* mig att yla av sorg. BergmanZ1

Er was een foto bij die mij *deed* janken van verdriet. BergmanN2

Slutsats

Efter en kort introduktion till korpuslingvistik har jag gett en beskrivning av arbetet med uppbyggnaden av den nederländsk-svenska parallellkorpus som jag använder mig av i min forskning om kausativa verb och verbkonstruktioner i nederländskan och svenskan. Till slut har jag illustrerat korpusarbetet i den kontrastiva analysen med utgångspunkt i en förundersökning vars målsättning var att kartlägga de olika kausativa verben som förekommer i konstruktioner med infinitivkomplement i båda språken.

Litteratur

Aarts, Jan. 1998. "Introduction." I: Johansson, Stig and Signe Oksefjell (red.). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi.

Aijmer, Karin & Altenberg, Bengt (red.). 1996. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London: Longman.

Aijmer, Karin, Altenberg, Bengt & Johansson, Mats. 1996. "Text-based contrastive studies in English. Presentation of a project." I: Aijmer, Karin, Altenberg, Bengt & Johansson, Mats (red.). *Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*. Lund: Lund University Press. s. 73-85.

Aijmer, Karin, Altenberg, Bengt & Johansson, Mats (red.). 1996. *Languages in Contrast Papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*. Lund: Lund University Press.

Altenberg,Bengt. 2002. Causative constructions in English and Swedish. A corpus-based contrastive study. I: Altenberg, Bengt; Granger, Sylviane (red.) *Lexis in Contrast*. Amsterdam: John Benjamins: s. 97-116.

Atkins, Sue, Clear Jeremy & Ostler Nicholas. 1992. "Corpus design criteria." I: *Literary and linguistic computing 7*. Oxford : Oxford university press, s. 1-16.

Barlow, Michael. 1999. "MonoConc 1.5 and ParaConc." I: *International Journal of Corpus Linguistics 4 (1)*. Amsterdam/Philadelphia: Benjamins, s. 319-327.

Church, Kenneth, Gale, William, Hanks, Patrick, Hindle, Donald, Bell Laboratories & Oxford University Press. 1991. "Using statistics in lexical analysis." In: Uri Zernik (red.). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale (New Jersey): Erlbaum,s.115-164.

Gale, W. A. & Church K. W. 1993. A Program for Aligning Sentences in Bilingual Corpora. I: *Computational Linguistics 19 (1)*, s. 75-101.

Cook V.J.. 1988. *Chomsky's Universal Grammar: An Introduction*. Oxford: Blackwell.

De Mönnink, Inge. 1997. "Using corpus and experimental data: a multi-method approach." I: M. Ljung (red.). *Studies in English Corpus Linguistics. Papers from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17).* Amsterdam: Rodopi. s. 227-244.

Francis, W.Nelson & Kucera, Henry. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

Garside, Roger, Leech, Geoffrey & Sampson, Geoffrey (red.). 1990. *The Computational Analysis of English : a corpus-based approach*. London: Longman.

Gellerstam, Martin. 1996. "Translations as a source for cross-linguistic studies." I: Aijmer, Karin, Altenberg, Bengt, & Johansson, Mats (red.). *Languages in Contrast.* Lund: Lund University Press, s. 53-62.

Hofland, Knut & Johansson, Stig. 1982. *Word frequencies in British and American English.* Bergen: The Norwegian Computing Centre for the Humanities.

Jespersen, Otto. 1938. *En sprogmands levned.* Köpenhamn: Gyldendal.

Johansson, Stig. 1975. *Papers in contrastive linguistics and language testing.* Lund: Gleerup.

Johansson, Stig. 1998. "On the role of corpora in cross-linguistic reasearch." I: Johansson, Stig & Oksefjell, Signe (red.). *Corpora and Cross-linguistic Research : Theory, Method, and Case Studies*. Amsterdam: Rodopi, s. 3-24.

Johansson, Stig. 2000. "Contrastive Linguistics and Corpora no.3." I: *SPRIKreports*
http://www.hf.uio.no/german/sprik/

Laureys, Godelieve & Rawoens, Gudrun. 2001. "Bilingval leksikografi. Nederlandsk-dansk ordbogsprojekt." I: *Nordiska studier i Lexikografi 5*. Göteborg, s. 185-201.

Lauridsen, K. M.. 1989. ”Tekstkorpora. Ny forskningsaktivitet ved Handelshøjskolen.” I: *Handelshøjskolen 50 år. Festskrift udgivet i anledning af Handelshøjskolens 50-års jubilæum 31. august 1989.* Aarhus: the Aarhus School of Business, s. 118-125.

Lauridsen, Karen M.. 1996. “Text corpora and contrastive linguistics: Which type of corpus for which type of analysis?” I: Aijmer, Karin, Altenberg, Bengt, & Johansson, Mats (red.). *Languages in Contrast.* Lund: Lund University Press, s. 63-71.

Leech, Geoffrey. 1996. “The state of the art in corpus linguistics.” I: Aijmer, Karin & Altenberg, Bengt (red.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London: Longman.

Leech, Geoffrey. 1991. “Corpora.” I: Malmkjaer, Kirsten (red.). *The linguistics encyclopedia*. London: Routledge. s. 73-80.

Leech, Geoffrey. 1992. “Corpora and theories of linguistic performance.” I: Svartvik, Jan (red.). *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, ["The State of the Art in Corpus Linguistics"], Stockholm, Sweden, August 4-8, 1991*. NY: Mouton de Gruyter, s. 105-122.

Leech, Geoffrey. 1966. *English in Advertising: A Linguistic Study of Advertising in Great Britain.* London: Longman.

McEnery, Tony & Wilson, Andrew. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, Vincent B. Y.. 1998. *Computer corpus lexicography*. Edinburgh: Edinburgh university press.

Ooi, Vincent B. Y.. 1994. “Corpus linguistics.” I: *SAAL Quarterly (Journal of the Singapore Association for Applied Linguistics) 28,* s.2-4.

Rawoens, Gudrun. 2002. A corpus-based Dutch-Swedish contrastive study of causative verbs and constructions. I: *Studies in Contrastive Linguistics. Proceedings Second International Contrastive Linguistics Conference, Santiago de Compostela, October 2001*, s. 863-873.

Sinclair, John. 1987. “Grammar in the dictionary.” I: Sinclair, John (red.). *Looking up: An Account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. London: Harper-Collins, s. 103-115.

Sinclair, John. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1997. “Corpus Evidence in Language Description.” I: Wichmann, Anne, Fligelstone, Steven, McEnery Tony & Knowles, Gerry (red.). *Teaching and Language Corpora*. Longman: London and New York. s. 27-39.

Stubbs, Mickael. 1995. “Collocations and semantic profiles. On the cause of the trouble with quantitative studies.” I: *Functions of language 2 (1)*. Amsterdam : Benjamins, s. 23-55.

Summers, Della. 1996. ”Computer lexicography: the importance of representativeness in relation to frequency.” I: Thomas Jenny & Short Mick (red.). *Using corpora for language research : studies in the honour of Geoffrey Leech*. London: Longman. s. 260-266.

Svartvik, Jan. 1992. “Corpus linguistics comes of age.” I: Svartvik, Jan (red.). *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, ["The State of the Art in Corpus Linguistics"], Stockholm, Sweden, August 4-8, 1991.* New York: Mouton de Gruyter, s. 7-13.

Woods, Anthony, Fletcher, Paul & Hughes, Arthur. 1986. *Statistics in language studies*. Cambridge: Cambridge university press.